Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer

Alabi, Rasheed Omobolaji

2020-04

Alabi , R O , Elmusrati , M , Sawazaki-Calone , I , Kowalski , L P , Haglund , C , Coletta , R D , Mäkitie , A A , Salo , T , Almangush , A & Leivo , I 2020 , ' Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer ' , International Journal of Medical Informatics , vol. 136 , 104068 . https://doi.org/10.1016/j.ijmedi

http://hdl.handle.net/10138/327123
https://doi.org/10.1016/j.ijmedinf.2019.104068

Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer

Rasheed Omobolaji Alabi M.Sc [a], Mohammed Elmusrati D.Sc [a], Iris Sawazaki-Calone DDS, PhD [b], Luiz Paulo Kowalski MD, PhD [c], Caj Haglund MD, PhD [d], Ricardo D. Coletta DDS, PhD [e], Antti A. Mäkitie MD, PhD [f], Tuula Salo DDS, PhD [g], Alhadi Almangush DDS, PhD [h*], Ilmo Leivo MD, PhD [i*]

[a] Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland.

[b] Oral Pathology and Oral Medicine, Dentistry School, Western Parana State University, Cascavel, PR, Brazil.

[c] Department of Head and Neck Surgery and Otorhinolaryngology, A.C. Camargo Cancer Center, São Paulo-SP, Brazil.

[d] Research Programs Unit, Translational Cancer Biology, University of Helsinki, Helsinki, Finland.
Department of Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland.

[e] Department of Oral Diagnosis, School of Dentistry, State University of Campinas, Piracicaba, São Paulo, Brazil.

[f] Department of Otorhinolaryngology – Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Helsinki, Finland.
Research Program in Systems Oncology, Faculty of Medicine, University of Helsinki, Helsinki, Finland.
Division of Ear, Nose and Throat Diseases, Department of Clinical Sciences, Intervention and Technology, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden.

[g] Department of Pathology, University of Helsinki, Helsinki, Finland.
Department of Oral and Maxillofacial Diseases, University of Helsinki, Helsinki, Finland.
Cancer and Translational Medicine Research Unit, Medical Research Center Oulu, University of Oulu and Oulu University Hospital, Oulu, Finland.

[h] Department of Pathology, University of Helsinki, Helsinki, Finland.
Institute of Biomedicine, Pathology, University of Turku, Turku, Finland
Faculty of Dentistry, University of Misurata, Misurata, Libya.

[i] University of Turku, Institute of Biomedicine, Pathology, Turku, Finland.

***The last two authors have equal contributions.**

Corresponding Author: Rasheed Omobolaji Alabi
Department of Industrial Digitalization, School of Technology and Innovations, University of Vaasa, Vaasa, Finland.
E-mail address: rasheed.alabi@student.uwasa.fi (or rasheed.alabii@yahoo.com), +358442056910

Disclosure: The authors declare no conflicts of interest.

Date: 28th May, 2019

Abstract

Background: The proper estimate of the risk of recurrences in early-stage oral tongue squamous cell carcinoma (OTSCC) is mandatory for individual treatment-decision making. However, this remains a challenge even for experienced multidisciplinary centers.

Objectives: We compared the performance of four machine learning (ML) algorithms for predicting the risk of locoregional recurrences in patients with OTSCC. These algorithms were Support Vector Machine (SVM), Naive Bayes (NB), Boosted Decision Tree (BDT), and Decision Forest (DF).

Materials and methods: The study cohort comprised 311 cases from the five University Hospitals in Finland and A.C. Camargo Cancer Center, São Paulo, Brazil. For comparison of the algorithms, we used the harmonic mean of precision and recall called F1 score, specificity, and accuracy values. These algorithms and their corresponding permutation feature importance (PFI) with the input parameters were externally tested on 59 new cases. Furthermore, we compared the performance of the algorithm that showed the highest prediction accuracy with the prognostic significance of depth of invasion (DOI).

Results: The results showed that the average specificity of all the algorithms was 71%. The SVM showed an accuracy of 68% and F1 score of 0.63, NB an accuracy of 70% and F1 score of 0.64, BDT an accuracy of 81% and F1 score of 0.78, and DF an accuracy of 78% and F1 score of 0.70. Additionally, these algorithms outperformed the DOI-based approach, which gave an accuracy of 63%. With PFI-analysis, there was no significant difference in the overall accuracies of three of the algorithms; PFI-BDT accuracy increased to 83.1%, PFI-DF increased to 80%, PFI-SVM decreased to 64.4%, while PFI-NB accuracy increased significantly to 81.4%. **Conclusions:** Our findings show that the best classification accuracy was achieved with the boosted decision tree algorithm. Additionally, these algorithms outperformed the DOI-based approach. Furthermore, with few parameters identified in the PFI analysis, ML technique still showed the ability to predict locoregional recurrence. The application of boosted decision tree machine learning algorithm can stratify OTSCC patients and thus aid in their individual treatment planning.

KEYWORDS: Artificial Intelligence; Oral tongue cancer; Machine Learning; Prediction

1. Introduction

Oral tongue squamous cell carcinoma (OTSCC) refers to squamous cell carcinoma that arises from the anterior two thirds of the tongue (also known as mobile tongue). It is usually reported as part of oral squamous cell carcinoma (OSCC), which includes all anatomical subsites of the oral cavity. A recent international study including 22 registries reported 89,212 incident cases of OTSCC and an increasing annual incidence [1], which has been confirmed by others [2]. The primary treatment of choice for OTSCC is surgical excision. However, even early-stage tumors may express a pattern of aggressive behavior [3,4]. Thus, OTSCC with aggressive behavior and those with advanced stage require multimodality treatment including neck dissection and adjuvant (chemo)radiotherapy. Therefore, it is important to precisely estimate the clinical behavior and outcome of OTSCC. Predicting the risk of recurrences is one of the important assessments for the clinician during treatment planning. More importantly, early diagnosis and predicting the risk of recurrences form a milestone in the management of OTSCC as the recent analysis of Finnish cases reported that about 67% of OTSCC cases were diagnosed at an early stage (I-II) [5]. With accurate and timely recurrence prediction, high-risk cases of OTSCC can be identified and multimodality treatment applied accordingly. In a large cohort of early OTSCC, about one fourth of cases (27.8%) developed a recurrence, and all of them might have benefitted from early prediction and corresponding treatment planning [6].

Many recent studies have examined the use of machine learning (ML) techniques for prognostication of different cancers [7,8]. Interestingly, predicting patient outcome by ML techniques has shown better accuracy than Cox regression [9]. This is why the use of ML has been in active research focus during recent years. For instance, ML techniques have been used to predict the outcome of various cancer types [10–12] and a web-based tool based on artificial neural network to predict outcome in cancer has been reported [13].

In this study, we examined four different ML algorithms, namely, support vector machine (SVM), naive Bayes (NB), boosted decision tree (BDT), and decision forest (DF) in terms of their performances to predict locoregional recurrence in OTSCC patients. Also, the predictive performance of a permutation feature importance (PFI) of these algorithms was evaluated. Many researchers have used this approach for comparing ML techniques for survival prediction in different malignancies like breast and lung cancers [14–17]. Tapak et al. examined six ML algorithms and two traditional methods for the prediction of breast cancer survival and metastasis [15]. In our study, we aimed to identify the best algorithm that would effectively classify patients as either low-risk or high-risk OTSCC recurrence. The algorithm with the overall best classification performance was further compared to a recently reported risk model based on the depth of invasion (DOI) [18]. This comparison was a result of the fact that DOI of 4mm or deeper has been considered to be a factor that accurately predicts locoregional recurrence [6]. Moreover, the recent American Joint Committee on Cancer (AJCC) 8th edition incorporated depth of invasion (DOI) into T-stage [19]. Similarly, the study by Almangush et al. suggested that DOI is one of the strongest pathological predictors for locoregional recurrence [6]. This suggestion is in agreement with reports by others [20,21].

We hypothesize that the application of the above-mentioned supervised learning classifiers may be used in the prediction of OTSCC locoregional recurrences and will thereby add value for the management of OTSCC.

2. Material and Methods

Patients: We used data from a study cohort comprising patients treated at the five Finnish University Hospitals of Helsinki, Oulu, Turku, Tampere, and Kuopio and at the A.C. Camargo Cancer Center, Sao Paulo, Brazil. This is a multicenter study from six institutions and data were provided for many cases as locoregional recurrences without specification. The

clinicopathologic characteristics of this cohort have been previously reported and summarized [22]. The primary treatment for all cases was surgical excision. In addition, some cases received neck dissection and/or adjuvant radiotherapy. The parameters included were age, gender, T-stage, WHO grade, tumor budding, depth of invasion, worst pattern of invasion (WPOI), lymphocytic host response (LHR), and perineural invasion (PNI) as shown in Table 1. Several studies have confirmed the prognostic importance of these variables [6,13,22–25]. Neck dissection and adjuvant radiotherapy were also included in the machine learning algorithms due to the impact of variation in the treatment modality that might influence the risk of recurrence. The use of patient samples and data inquiry were approved by the Hospital Research Ethics Committees of all individual hospitals, by the Finnish National Supervisory Authority for Welfare and Health (VALVIRA) and by the Brazilian Human Research Ethics Committee.

2.1. The classification algorithms examined

The algorithms considered in this study are basic and have been commonly used in other cancer studies [14–18].

2.1.1. Support vector machine (SVM) is an elegant and powerful ML technique extensively used for both classification and regression problems [26]. This is due to its ability to classify non-linearly separable patterns by projecting the original features into a higher dimensional space (hyperplane) [27].

2.1.2. Naive Bayes (NB) is known as Bayes point machine in the Azure ML studio and it is based on the generally-known Bayes theorem [26,27]. The algorithm operates by learning and estimating the prior probability of belonging to each class using the training data. [27,28].

2.1.3. Boosted Decision Tree (BDT) with gradient boosting machine was the subtype of BDT used in this study. It is an ensemble learning method where the second tree corrects the errors of the first tree, the third tree corrects the errors in the second trees, the fourth tree corrects the errors in the third trees, etc. Predictions are based on the entire ensemble of trees [27,28].

2.1.4. Decision Forest (DF) relies on the combination of multiple related models to get better results and a more generalized model. Therefore, it works by using a bootstrapped sample of data to build each tree where only a proportion of the variable set is considered for each tree. Each tree in the decision forest outputs a frequency histogram of labels that is non-normalized. These frequency histograms were aggregated in the process that sums these histograms and then normalizes the results to get the probabilities for each label [27].

2.1.5. Permutation Feature Importance (PFI) is a model-agnostic ranker feature ranker that computes the scores for each of the variables contained in a dataset. It basically examines the contribution of each feature to the overall predictive performance of the algorithm [27].

2.2. Evaluation of the performance of the algorithms

The performance metrics were aimed to evaluate how the algorithms performed [29–31]. Most of these metrics have been previously used in other studies [15,32]. However, in addition to accuracy, only two (F1 score and specificity) of these statistical measures that are medically more relevant in the clinic, were discussed in the current study.

3. The training-validation phase for the algorithms in Microsoft Azure for prediction of recurrence

Microsoft Azure Machine Learning Studio (Azure ML 2019) was used in this study [27]. The input parameters were age, gender, stage, grade, tumor budding, depth of invasion (DOI), worst

pattern of invasion (WPOI), lymphocytic host response (LHR), perineural invasion (PNI) and treatment given, while the target output was locoregional recurrence. Disease-free survival (DFS) time of the cases ranged from 1 to 267 months. Specifically, the DFS in cases with recurrence varied between 1 and 120 months. Firstly, a potential class imbalance with respect to the number of patients who experienced a tumor recurrence in the target class (locoregional recurrence) was handled by up-sampling in order to balance the classes used in the training. Synthetic minority oversampling technique (SMOTE) [33] offers a better way to handle imbalance than simply duplicating existing cases. The dataset and the corresponding samples are therefore more general [33]. The dataset was divided into two sets of training and validation. Due to the relatively limited amount of data, a 5-fold cross validation was used with 50% training and 50% validation {50:50} percentage splitting sets [15]. Each of the algorithms of interest was then configured as shown in Figure 1 [27,28]. After training, the algorithms were evaluated for the various quality metrics (Table 3).

Furthermore, these algorithms were further tested with new cases (**Section 3.1**). The result obtained from this approach was considered as the gold standard in this study as it gives an account of how the algorithm is expected to predict in reality. Also, it addresses any concerns about the generalizability of the trained models. In addition, the contribution of each of the input variables on the predictive ability of each model was examined using permutation feature importance (PFI) analysis. Their contributions were given in the form of PFI-performance scores. To avoid bias in the algorithm, disease-free survival and treatment were removed from the PFI analysis that was aimed to examine the predictive ability of each variables. The input features with positive scores were selected. Also, only one of the inputs was selected when two or more inputs give the same negative score. The variables with least scores were not selected. These selected variables were used to train the algorithms. The given accuracies in the PFI analysis were compared with the accuracies obtained without PFI. Similarly, the PFI-based algorithms were tested with new cases.

3.1 Testing performance of the model with new cases: In this phase, the trained algorithms were tested with 59 new cohort cases that were not included in the training or in the validation sets (Figure 1). These new independent data were obtained from a cancer center in Brazil. The results are presented in Table 4. The PFI-based models were also tested with these new cases as presented in Table 5.

3.2 Comparison with the depth of invasion (DOI): The algorithm that showed the highest overall accuracy when tested with these new external cases (section 3.1) was also compared with the depth of invasion (DOI) based model as shown in Figure 3**.**

4. Results

4.1 Data Description

The study cohort included 311 patients with cT1-T2cN0M0 OTSCC; 165 men and 146 women, resulting in a male-to-female range of 1.1:1. Out of these 311 cases, 57 cases had missing details about any postoperative treatment information. Therefore, these cases were excluded and the machine learning training was performed with 254 cases. These cases included 141 men and 113 women with the mean age at diagnosis was 61.51 (SD $\pm$ 14.81: range 10-95) and the median age was 62.0 years. The distribution according to tumor diameter showed that 100 patients had stage T1 and 154 stage T2. The histopathologic parameters are briefly summarized in Table 2**.** In terms of the treatment, 157 patients had surgery alone while 97 had adjuvant (chemo)radiotherapy (92 radiotherapy and 5 chemoradiotherapy) . Similarly, 185 had neck dissection while 69 had no neck dissection performed. Thus, from the 185 patients who had neck dissection, 43% were exposed to adjuvant radiotherapy while 57% had only surgery as single-modality treatment. Similarly, out of the 69 cases who had no dissection performed, 25% were exposed to adjuvant radiotherapy while 75% had only surgery.

The number of patients with disease recurrences was 68 (26.8%). While the disease-free survival (DFS) time ranged from 1 to 267 months, the DFS time for cases with a locoregional recurrence was between 1 to 120 months. Overall, 89.6% of the recurrences occurred in the first 2 years, while 10.45% recurrence was recorded after 2 years. The mean follow-up time was 75 months (SD $\pm$ 64.6; range 1 - 258 months) and the median was 60 months. Similarly, for the 59 new OSCC cases used for external testing, DFS time varied between 1 to 146 months. Also, 74% had a recurrence in the first year, 16% after the first and before end of second year, and 10% of the patients recurred after the second year. The mean age in this external validation cohort was 56.2 years (range, 31-84 years). All these new cases had neck dissection, where 34 cases had surgery alone while 25 had adjuvant (chemo)radiotherapy (22 radiotherapy and 3 chemoradiotherapy). The DOI model performance in terms of accuracy in the training set was 47.2% and the overall accuracy in the new cohorts used for external validation was 63%.

Performance metrics for the algorithms

The distribution of true and false positives, true and false negatives, and other performance metrics for the algorithms in the training phase are given in Figure 2a and Table 3, respectively. During the training phase, decision forest showed the highest accuracy while naive Bayes and decision forest showed the best area under receiving operating characteristic (AUC of ROC). When these algorithms were tested on the 59 new external cases from the cancer center in Brazil, the average specificity of all the algorithms was 71%. The tested algorithms i.e. support vector machine, naive Bayes, decision forest, and boosted decision tree gave an overall accuracy of 68%, 70%, 78% and 81%, respectively. The details of the performance of parameters with this new cohorts are given in Table 4. Considering the harmonic mean of precision and recall, that is, F1 score, the support vector machine, naive Bayes, decision forest, and boosted decision tree gave 0.63, 0.64, 0.70 and 0.78, respectively. Therefore, the best

overall classification performance to predict recurrence was achieved with the boosted decision tree algorithm. Comparison of the boosted decision tree algorithm and the DOI model is shown in Figure 3; the DOI model showed an accuracy of 63% where 54.1% of the patients would be observed, thereby not subjected to adjuvant therapy or elective neck dissection (END). The boosted decision tree on the other hand showed 81% overall accuracy where 21.1% of the patients would have been observed and not subjected to END. Similarly, about half (49.5%) of the patients were correctly identified as having OTSCC recurrence using the DOI model. Boosted decision tree machine learning technique correctly identified 78.9% as having OTSCC recurrence as shown in Figure 3. Thus, each of these algorithms performed significantly better than the DOI-based model.

The results of the permutation feature importance (PFI) analyses are given in Table 5. The PFI scores were calculated for each feature independently. A zero score is returned when there is no difference in the performance metrics before and after PFI of that feature. Similarly, a negative score is returned when a random PFI of that feature produced a higher accuracy and lower error (better performance metrics) compared to the performance before PFI was applied. Moreover, a higher importance score (positive) gives an indication of the contribution of that feature to the predictive ability of the model. The PFI of boosted decision tree (PFI-BDT) showed the highest accuracy (83.1%). Also, it was observed that the accuracy of BDT increased from 81.0% to 83.1% and DF increased from 78% to 80%, while SVM showed a reduction in accuracy from 68% to 64.4% in the PFI analysis. Interestingly, the accuracy of NB increased significantly from 70.0% to 81.4% in the permutation feature importance fitting. The ranking of the scores of the features is as shown in Table 5.

5. Discussion

The present study compared the performance of ML algorithms to stratify patients with OTSCC into low or high-recurrence risk group. In this regard, four ML algorithms, namely, boosted decision tree, naive Bayes, support vector machine, and decision forest were examined. We

found that the performance of these techniques was higher than that of depth of invasion (DOI) based approach. Our multicenter cohort of cases is one of the largest published series. Majority of the previous publications including hundreds of cases have mixed early-stage cases with those with advanced stage, and/or have mixed different subsites of the oral cavity (e.g. oral tongue with floor of mouth and retromolar region). Therefore, heterogeneity of such series makes it difficult to identify robust prognostic markers. The advantage of our homogenous cohort (only early stage and only oral tongue) allows for reaching definitive conclusions that can be considered to be applied in daily practice.

Although significant progress has been made in early diagnostics, treatment strategies and prevention of OTSCC in recent years, the prognosis of OTSCC is poor due to aggressive local invasion and metastasis, leading to recurrence. The mortality rates in cases with recurrence has been reported to be very high [34]. When recurrence is diagnosed earlier, the mortality rates have been reported to decrease [35,36]. The reported rates of recurrence in oral squamous cell carcinoma range from 6.9 % to 37.4% of patients [37,38]. This is in accordance with the 26.8% locoregional recurrence rate within the dataset used in this study. Improved prediction of locoregional recurrences in early-stage OTSCC can lead to an adjusted, patient-oriented follow-up program. For example, based on prediction of the patient as a high-risk case a customized surveillance could be organized instead of the general follow-up program.

Abundant studies exist that have considered DOI as a strong histologic feature that correlates with locoregional recurrence. The machine learning algorithms examined in this study, however, outperformed the power of prediction of locoregional recurrence based on DOI. However, it will offer a better approach with significant accuracy in stratifying the patients as carrying a high- or low-risk for recurrence. Therefore, it seems obvious, that the described challenge in the treatment-decision making would be successfully addressed by the machine learning model due to increased specificity, F1 score and overall accuracies of the ML

algorithms. Thus, this study has potentially high impact to clinicians in the management of early OTSCC.

With regards to the performance metrics examined, F1 score used as the benchmark to choose the best algorithm as it finds the optimal blend between two other performance metrics (precision and recall). As shown in Table 4, the F1 score for the boosted decision tree algorithm showed to be very good at stratifying the patients as having either low-risk or high-risk of recurrence of OTSCC. This justifies why boosted decision tree was compared to the DOI as shown in Figure 3 [18]. It is important to note that the support vector machine showed promising evaluation performance metrics in the training phase. This is due to the fact that it is an empirical risk minimizer algorithm. Hence, it is not usually prone to overfitting related issue as it avoids the danger of getting trapped into local minima [39]. However, the ensemble algorithms performed better than the support vector machine because they were able to create a fleet of algorithms with relatively similar bias and subsequently combining their outputs to reduce variance.

Furthermore, a major challenge in the treatment of patients with early OTSCC is in finding the right parameters that predict prognosis and help to properly identify patients at high risk of locoregional recurrences. This would carry the potential to minimize the incidence treatment failure of patients with OTSCC [35]. With the PFI-analyses, the exact contribution of each parameter to the predictive ability of the machine learning algorithms was known. Interestingly, there was no significant difference in the overall accuracies achieved in the ensemble methods (decision forest and boosted decision tree) with reduced parameters identified in the PFI analyses compared to the algorithms without PFI. Therefore, the cost and resources associated with getting numerous parameters can be properly managed. Also, the time taken to properly prepare an individualized treatment plan for the patients can be improved. This is because a few but important features that are needed for the ML algorithms were identified in the PFI analysis while producing the same range of prediction accuracies.

Thus, predicting recurrence with such accuracy as shown in this study would be crucial to the clinicians in terms of management decisions.

Numerous studies have compared the performance of various machine learning classifiers to predict an outcome of interest in cancer. Tapak et al. compared various machine learning classifiers in series of 550 breast cancer patients, and found that the support vector machine predicted survival better than other classifiers [15]. Similarly, the study by Tseng et al. compared decision tree ML technique with a traditional statistical model such as logistic regression in series of 673 oral cancer patients and the decision tree was found to perform better [40]. De Melo et al. used decision tree to evaluate the quality of life among patients with head and neck cancer [41]. Similarly, Sumbaly et al. used the decision tree in the diagnosis of breast cancer [42]. The decision forest also produced the highest prognostic performance when compared with other machine algorithms by Zhang et al. for the radiomics-based prediction of failure in advanced nasopharyngeal carcinoma [43].

In conclusion, this study investigated four different ML algorithms and found that the boosted decision tree algorithm showed the best overall performance accuracy. Due to the sensitive nature of the application of machine learning in medicine, it is important for these algorithms to produce very high accuracies. In this study, the ensemble algorithms such as the boosted decision tree and the decision forest algorithms performed better than non-ensemble algorithms such as support vector machine, naive Bayes and a method based on depth of invasion. Therefore, the ensemble machine algorithms should be considered in medical applications. Presently, it is challenging for clinicians to assess the outcomes of clinical early-stage oral cancer. For the clinicians, knowledge of potential locoregional prediction to stratify the patients into low-risk or high-risk groups using machine learning applications can help to guide clinical practice. Patients can be counseled accordingly with realistic expectations and clinicians can be guided in making informed decisions. Furthermore, this contributes to the individual data regarding patient and tumor-related factors and thereby helps the clinician in

planning the optimal patient-specific treatment and follow-up (post-operative adjuvant treatment). For instance, high-risk patients might benefit from adjuvant oncological therapy after surgery. Future research should consider including other prognostic parameters as inputs for the selected algorithms. In terms of the limitation of this study, we are limited by the number of available cases as this was a retrospective study of five teaching hospitals in Finland and one in Brazil. Also, the external data used to test the performance of the algorithms were relatively limited. Therefore, with larger external data, the performance of the algorithms could be improved.

Authors Contribution

Institutional Coordinators: Salo T, Coletta RD, Kowalski LP, Leivo I, Mäkitie AA, Haglund C. **Study concepts and study design:** Alabi RO, Elmusrati M, Almangush A, Coletta RD, Salo T, Leivo I. **Data acquisition and quality control of data:** Sawazaki-Calone I, Kowalski LP, Leivo I. **Data analysis and interpretation:** Alabi RO, Elmusrati M, Almangush A, Sawazaki-Calone I, Mäkitie AA, Salo T, Leivo I. **Manuscript preparation:** Alabi RO, Elmusrati M, Almangush A, Mäkitie AA, Coletta RD. **Manuscript review:** Mäkitie AA, Leivo I, Salo T, Kowalski LP, Sawazaki-Calone I. **Manuscript editing:** Salo T, Leivo I, Mäkitie AA, Haglund C. All authors approved the final manuscript for submission.

Summary points

What was already known on the topic:

- There are few published studies on the comparison of machine learning techniques to predict locoregional recurrence of oral tongue squamous cell carcinoma (OTSCC).

- Accuracy value is the most considered performance metrics to choosing the machine learning technique for prediction.

What knowledge this study adds:

- To the best of our knowledge, this is the first study that analyzed more than three machine learning techniques to predict risk of locoregional recurrence in oral tongue squamous cell carcinoma (OTSCC) as low-risk or high-risk.
- It is important to consider other performance metrics such as specificity and F1 score (weighted average of precision and recall) in medical applications.
- The permutation importance feature (PFI) algorithm to extract important features does not correspond to better overall prediction and does not necessarily perform better than the ensemble algorithms.
- The application of these supervised learning techniques to stratify the patients as having low-risk or high-risk for the recurrence of OTSCC may be useful for effective cancer management.

Acknowledgment

The School of Technology and Innovations, University of Vaasa Scholarship Fund. The Helsinki University Hospital Research Fund.

Figure Legend

Figure 1. The training process in azure machine learning studio**.**

Figure 2. The classification results of the four basic parameters for each algorithm in the training and also for PFI analysis.

(TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative, BDT: Boosted Decision Tree, SVM: Support Vector Machine, NB: Naive Bayes, and DF: Decision Forest).

Figure 3. The comparison of the boosted decision tree algorithm to the depth of Invasion model [18]

References

[1] J.H. Ng, N.G. Iyer, M.-H. Tan, G. Edgren, Changing epidemiology of oral squamous cell carcinoma of the tongue: A global study: Changing epidemiology of tongue cancer, Head Neck. 39 (2017) 297–304. doi:10.1002/hed.24589.

[2] J.E. Tota, W.F. Anderson, C. Coffey, J. Califano, W. Cozen, R.L. Ferris, M. St. John, E.E.W. Cohen, A.K. Chaturvedi, Rising incidence of oral tongue cancer among white men and women in the United States, 1973–2012, Oral Oncology. 67 (2017) 146–152. doi:10.1016/j.oraloncology.2017.02.019.

[3] K. Rusthoven, A. Ballonoff, D. Raben, C. Chen, Poor prognosis in patients with stage I and II oral tongue squamous cell carcinoma, Cancer. 112 (2008) 345–351. doi:10.1002/cncr.23183.

[4] I.O. Bello, Y. Soini, T. Salo, Prognostic evaluation of oral tongue cancer: Means, markers and perspectives (I), Oral Oncology. 46 (2010) 630–635. doi:10.1016/j.oraloncology.2010.06.006.

[5] R. Mroueh, A. Haapaniemi, R. Grénman, J. Laranne, M. Pukkila, A. Almangush, T. Salo, A. Mäkitie, Improved outcomes with oral tongue squamous cell carcinoma in Finland: Oral tongue carcinoma in Finland, Head Neck. 39 (2017) 1306–1312. doi:10.1002/hed.24744.

[6] A. Almangush, I.O. Bello, R.D. Coletta, A.A. Mäkitie, L.K. Mäkinen, J.H. Kauppila, M. Pukkila, J. Hagström, J. Laranne, Y. Soini, V.-M. Kosma, P. Koivunen, N. Kelner, L.P. Kowalski, R. Grénman, I. Leivo, E. Läärä, T. Salo, For early-stage oral tongue cancer, depth of invasion and worst pattern of invasion are the strongest pathological predictors for locoregional recurrence and mortality, Virchows Arch. 467 (2015) 39–46. doi:10.1007/s00428-015-1758-z.

[7] S. Anand, K. Rajesh, Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm, International Journal of Advanced Research in Computer and Communication Engineering. 1 (2012) 72–77.

[8] B. Zheng, S.W. Yoon, S.S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, Expert Systems with Applications. 41 (2014) 1476–1482. doi:10.1016/j.eswa.2013.08.044.

[9] L. Zhu, W. Luo, M. Su, H. Wei, J. Wei, X. Zhang, C. Zou, Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients, Biomedical Reports. 1 (2013) 757–760. doi:10.3892/br.2013.140.

[10] D. Delen, N. Patil, Knowledge Extraction from Prostate Cancer Data, in: Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), 2006: pp. 92b–92b. doi:10.1109/HICSS.2006.240.

[11] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine. 34 (2005) 113–127. doi:10.1016/j.artmed.2004.07.002.

[12] D. Delen, Analysis of cancer data: a data mining approach, Expert Systems. (2009). https://onlinelibrary.wiley.com/doi/full/10.1111/j.1468-0394.2008.00480.x (accessed September 5, 2019).
[13] R.O. Alabi, M. Elmusrati, I. Sawazaki-Calone, L.P. Kowalski, C. Haglund, R.D. Coletta, A.A. Mäkitie, T. Salo, I. Leivo, A. Almangush, Machine learning application for prediction of locoregional recurrences in early oral tongue cancer: a Web-based prognostic tool, Virchows Arch. (2019). doi:10.1007/s00428-019-02642-5.
[14] C.-M. Chao, Y.-W. Yu, B.-W. Cheng, Y.-L. Kuo, Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree, J Med Syst. 38 (2014) 106. doi:10.1007/s10916-014-0106-1.
[15] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, J. Poorolajal, Prediction of survival and metastasis in breast cancer patients using machine learning classifiers, Clinical Epidemiology and Global Health. (2018). doi:10.1016/j.cegh.2018.10.003.
[16] M. Montazeri, M. Montazeri, M. Montazeri, A. Beigzadeh, Machine learning models in breast cancer survival prediction, THC. 24 (2016) 31–42. doi:10.3233/THC-151071.
[17] C.M. Lynch, B. Abdollahi, J.D. Fuqua, A.R. de Carlo, J.A. Bartholomai, R.N. Balgemann, V.H. van Berkel, H.B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques, International Journal of Medical Informatics. 108 (2017) 1–8. doi:10.1016/j.ijmedinf.2017.09.013.
[18] A.M. Bur, A. Holcomb, S. Goodwin, J. Woodroof, O. Karadaghy, Y. Shnayder, K. Kakarala, J. Brant, M. Shew, Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma, Oral Oncology. 92 (2019) 20–25. doi:10.1016/j.oraloncology.2019.03.011.
[19] W.M. Lydiatt, S.G. Patel, B. O'Sullivan, M.S. Brandwein, J.A. Ridge, J.C. Migliacci, A.M. Loomis, J.P. Shah, Head and neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual: Head and Neck Cancers-Major 8th Edition Changes, CA: A Cancer Journal for Clinicians. 67 (2017) 122–137. doi:10.3322/caac.21389.
[20] M.J. Lin, A. Guiney, C.E. Iseli, M. Buchanan, T.A. Iseli, Prophylactic neck dissection in early oral tongue squamous cell carcinoma 2.1 to 4.0 mm depth, Otolaryngol Head Neck Surg. 144 (2011) 542–548. doi:10.1177/0194599810394988.
[21] P. O-charoenrat, G. Pillai, S. Patel, C. Fisher, D. Archer, S. Eccles, P. Rhys-Evans, Tumour thickness predicts cervical nodal metastases and survival in early oral tongue cancer, Oral Oncol. 39 (2003) 386–390.
[22] A. Almangush, R.D. Coletta, I.O. Bello, C. Bitu, H. Keski-Säntti, L.K. Mäkinen, J.H. Kauppila, M. Pukkila, J. Hagström, J. Laranne, S. Tommola, Y. Soini, V.-M. Kosma, P. Koivunen, L.P. Kowalski, P. Nieminen, R. Grénman, I. Leivo, T. Salo, A simple novel prognostic model for early stage oral tongue cancer, International Journal of Oral and Maxillofacial Surgery. 44 (2015) 143–150. doi:10.1016/j.ijom.2014.10.004.
[23] N. Yamakawa, T. Kirita, M. Umeda, S. Yanamoto, Y. Ota, M. Otsuru, M. Okura, H. Kurita, S. Yamada, T. Hasegawa, T. Aikawa, T. Komori, M. Ueda, Japan Oral Oncology Group (JOOG), Tumor budding and adjacent tissue at the invasive front correlate with delayed neck metastasis in clinical early-stage tongue squamous cell carcinoma, J Surg Oncol. (2018) jso.25334. doi:10.1002/jso.25334.
[24] X. Yang, X. Tian, K. Wu, W. Liu, S. Li, Z. Zhang, C. Zhang, Prognostic impact of perineural invasion in early stage oral tongue squamous cell carcinoma: Results from a prospective randomized trial, Surgical Oncology. 27 (2018) 123–128. doi:10.1016/j.suronc.2018.02.005.
[25] A. Arora, N. Husain, A. Bansal, A. Neyaz, R. Jaiswal, K. Jain, A. Chaturvedi, N. Anand, K. Malhotra, S. Shukla, Development of a New Outcome Prediction Model in

Early-stage Squamous Cell Carcinoma of the Oral Cavity Based on Histopathologic Parameters With Multivariate Analysis: The Aditi-Nuzhat Lymph-node Prediction Score (ANLPS) System, The American Journal of Surgical Pathology. 41 (2017) 950–960. doi:10.1097/PAS.0000000000000843.
[26] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Elsevier, 2011. doi:10.1016/C2009-0-19715-5.
[27] Microsoft Azure Machine Learning Studio, Azure Machine Learning Studio: In Documentation, (2018).
[28] R. Barga, V. Fontama, W.-H. Tok, Predictive analytics with Microsoft Azure Machine Learning, Second edition, Apress, Berkeley, CA, 2015.
[29] P.A. Flach, The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics, in: ICML, 2003.
[30] J. Fürnkranz, P.A. Flach, ROC 'n' Rule Learning—Towards a Better Understanding of Covering Algorithms, Mach Learn. 58 (2005) 39–77. doi:10.1007/s10994-005-5011-x.
[31] D.M.W. Powers, Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation, J of Mach Lear Tech. 2 (2011) 37–63.
[32] A.K. Dwivedi, Analysis of computational intelligence techniques for diabetes mellitus prediction, Neural Computing and Applications. 30 (2018) 3837–3845.
[33] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC Bioinformatics. 14 (2013) 106. doi:10.1186/1471-2105-14-106.
[34] J. Berdugo, L.D.R. Thompson, B. Purgina, C.D. Sturgis, M. Tuluc, R. Seethala, S.I. Chiosea, Measuring Depth of Invasion in Early Squamous Cell Carcinoma of the Oral Tongue: Positive Deep Margin, Extratumoral Perineural Invasion, and Other Challenges, Head and Neck Pathol. 13 (2019) 154–161. doi:10.1007/s12105-018-0925-3.
[35] A.-F. Safi, M. Kauke, A. Grandoch, H.-J. Nickenig, J.E. Zöller, M. Kreppel, Analysis of clinicopathological risk factors for locoregional recurrence of oral squamous cell carcinoma – Retrospective analysis of 517 patients, Journal of Cranio-Maxillofacial Surgery. 45 (2017) 1749–1753. doi:10.1016/j.jcms.2017.07.012.
[36] I. Vázquez-Mahía, J. Seoane, P. Varela-Centelles, I. Tomás, A.Á. García, J.L. López Cedrún, Predictors for Tumor Recurrence After Primary Definitive Surgery for Oral Cancer, Journal of Oral and Maxillofacial Surgery. 70 (2012) 1724–1732. doi:10.1016/j.joms.2011.06.228.
[37] M.A. Ermer, K. Kirsch, G. Bittermann, T. Fretwurst, K. Vach, M.C. Metzger, Recurrence rate and shift in histopathological differentiation of oral squamous cell carcinoma – A long-term retrospective study over a period of 13.5 years, Journal of Cranio-Maxillofacial Surgery. 43 (2015) 1309–1313. doi:10.1016/j.jcms.2015.05.011.
[38] D.R. Camisasca, M.A.N.C. Silami, J. Honorato, F.L. Dias, P.A.S. de Faria, S. de Q.C. Lourenço, Oral Squamous Cell Carcinoma: Clinicopathological Features in Patients with and without Recurrence, ORL. 73 (2011) 170–176. doi:10.1159/000328340.
[39] G. Levitin, Computational intelligence in reliability engineering, Springer, Berlin, 2007. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=186979 (accessed September 6, 2019).
[40] W.-T. Tseng, W.-F. Chiang, S.-Y. Liu, J. Roan, C.-N. Lin, The Application of Data Mining Techniques to Oral Cancer Prognosis, J Med Syst. 39 (2015) 59. doi:10.1007/s10916-015-0241-3.
[41] N.B. de Melo, Í. de M. Bernardino, D.P. de Melo, D.Q.C. Gomes, P.M. Bento, Head and neck cancer, quality of life, and determinant factors: a novel approach using decision tree analysis, Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology. 126 (2018) 486–493. doi:10.1016/j.oooo.2018.07.055.

[42] R. Sumbaly, N. Vishnusri, S. Jeyalatha, Diagnosis of Breast Cancer using Decision Tree Data Mining Technique, International Journal of Computer Applications. 98 (2014) 16–24.
[43] B. Zhang, X. He, F. Ouyang, D. Gu, Y. Dong, L. Zhang, X. Mo, W. Huang, J. Tian, S. Zhang, Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma, Cancer Letters. 403 (2017) 21–27. doi:10.1016/j.canlet.2017.06.004.

Table 1. The parameters contained in the dataset and their respective descriptors.

Number	Parameters	Description	Type
1	Age	Age at the time of diagnosis	Discrete
2	Gender	The sexual orientation of the patient	Categorical 1 = Male; 2 = Female
3	T-stage	T stage describing tumor size	Categorical 1 = T1; 2 = T2.
4	WHO Grade	Histopathologic grading according to World Health Organization (WHO) criteria	Categorical 1 = Grade I; 2 = Grade II; 3 = Grade III
5	Tumor budding	Tumor budding is defined as the presence of single cells or small clusters of cancer cells detached from the main tumor mass	Categorical 0 = No budding; 1 < 5 buds; 2 for $\geq$ 5 buds.
6	Tumor depth	This is the measure of tumor depth of invasion. It was measured in millimetres (mm)	Categorical 1 for < 4mm, 2 for $\geq$ 4mm
7	WPOI	Worst pattern of invasion	Categorical Value of 0 for WPOI type 1 to 3; Value of 1 for WPOI type 4; Value of 3 for WPOI type 5.
8	LHR	Lymphocytic host response	Categorical Value of 0 for LHR type 1; Value of 1 for LHR type 2; Value of 3 for LHR type 3.
9	PNI	Perineural invasion	Categorical 0 = Absent; 1 = Present
10	Treatment	This indicates the type of treatment offered for the patient. It could either be surgery alone or adjuvant (chemo)radiotherapy in addition to the surgery	Categorical 0 = Surgery alone 1 = Surgery + Adjuvant (chemo)radiotherapy
11	Neck treatment	This variable indicates whether neck dissection was performed or not	Categorical 0 = No neck dissection 1 = Neck dissection performed.
12	Recurrence*	The occurrence of disease after treatment	Categorical 0 = Low-Risk; 1 = High-Risk

* Recurrence was considered as the output/target label. The disease-survival (DFS) ranges from 1 to 267 months while DFS for locoregional recurrence patient ranges from 1 to 120 months.

Table 2: Summary of histopathologic parameters included for the machine learning training.

Variable	Category (Definition)	Number
WHO grade		
	Grade I (Well-differentiated tumor)	78
	Grade II (Moderately-differentiated tumor)	103
	Grade III (Poorly-differentiated tumor)	73
Tumor budding		
	None (There is no tumor budding)	93
	Low (Tumor has less than five buds)	85
	High (Tumor has five buds or more at the invasive front)	76
Depth of invasion		
	Superficial (Tumor < 4 mm in depth)	96
	Deep (Tumor ≥ 4 mm in depth)	158
Worst pattern of invasion (WPOI)		
	Type 1 (Pushing border) Type 2 (Finger-like growth) Type 3 (Large tumor islands)	64
	Type 4(Small tumor islands of ≤ 15 cancer cells)	158
	Type 5 (Tumor satellites)	32
Lymphocytic host response (LHR)		
	Type 1 (Strong)	36
	Type 2 (Intermediate)	88
	Type 3 (Weak)	130
Perineural invasion (PNI)		
	Absent (PNI was not observed)	223
	Present (PNI was observed)	31

Table 3. The overall performance metrics of the classifiers in the training phase

50% Training and 50% Testing Cross Validation Scheme									
Algorithm	**Sensitivity**	**Specificity**	**Precision**	**NPV**	**LR $^{+}$**	**LR $^{-}$**	**F1 Score**	**AUC**	**Accuracy %**
NB	0.67	0.81	0.77	0.92	3.53	0.41	0.66	0.89	80.0
SVM	0.94	0.79	0.59	0.97	4.48	0.08	0.73	0.88	82.7
DF	0.77	0.86	0.65	0.92	5.50	0.27	0.71	0.89	84.0
BDT	0.68	0.87	0.62	0.89	5.23	0.37	0.65	0.82	82.0
PFI-NB	0.77	0.83	0.59	0.92	4.53	0.28	0.67	0.89	81.0
PFI-SVM	0.87	0.72	0.50	0.95	3.11	0.18	0.64	0.87	76.0
PFI-DF	0.77	0.83	0.60	0.92	4.53	0.28	0.68	0.85	82.0
PFI-BDT	0.65	0.85	0.59	0.88	4.33	0.41	0.62	0.84	80.0

BDT = Boosted Decision Tree, SVM = Support Vector Machine, BPM = Bayes Point Machine, DF = Decision Forest, Precision (PPV = Predictive positive value), NPV = Negative predictive value, LR^{+} = Positive likelihood ratio and LR^{-} = Negative likelihood ratio, Sensitivity (recall), Area under receiving operating characteristics curve (AUC), and CDE = Custom Designed Ensemble.

Table 4. The performance of the algorithms with external cases.

Parameter	SVM	NB	BDT	DF
True Positive (TP)	16	16	15	15
False Positive (FP)	16	15	07	09
True Negative (TN)	24	25	33	31
False Negative (FN)	03	03	04	04
Sensitivity	0.84	0.84	0.79	0.79
Specificity	0.60	0.63	0.83	0.78
Precision (PPV)	0.50	0.52	0.76	0.63
NPV	0.89	0.89	0.89	0.89
LR^{+}	2.10	2.27	4.65	3.59
LR^{-}	0.27	0.25	0.25	0.27
F1 Score	0.63	0.64	0.78	0.70
Accuracy	**68%**	**70%**	**81%**	**78%**

Table 5. Permutation Feature Importance (PFI) of the algorithms.

PFI-DF		PFI-BDT		PFI-SVM		PFI-NB	
Features	**Scores**	**Features**	**Scores**	**Features**	**Scores**	**Features**	**Scores**
PNI	0.0078	Age	0.0315	Gender	0.0079	Age	0.0079
Depth	0.0000	Depth	0.0236	Stage	0.0079	Gender	0.0079
Tumor Budding	0.0158*	WPOI	0.0236	Tumor Budding	0.0079	Stage	0.0079
Stage	0.0315*	PNI	0.0079	Depth	0.0079	Depth	0.0079
LHR	0.0315*	Tumor Budding	0.0000	LHR	0.0079	Grade	0.0000
Gender	0.0394*	LHR	0.0079*	PNI	0.0079	Tumor Budding	0.0079*
Grade	0.0394*	Stage	0.0158*	Age	0.0000	LHR	0.0079*
WPOI	0.0472*	Grade	0.0158*	Grade	0.0000	PNI	0.0079*
Age	0.0551*	Gender	0.0236*	WPOI	0.0000	WPOI	0.0236*
Accuracy (External Testing)	80.0%	**Accuracy (External Testing)**	83.1%	**Accuracy (External Testing)**	64.4%	**Accuracy (External Testing)**	81.4%

* Negative score. DF : Decision Forest, BDT: Boosted Decision Forest, SVM: Support Vector Machine, NB: Naive Bayes. WPOI: Worst Pattern of Invasion, PNI: Perineural Invasion, LHR: Lymphocytic host response.

Figures

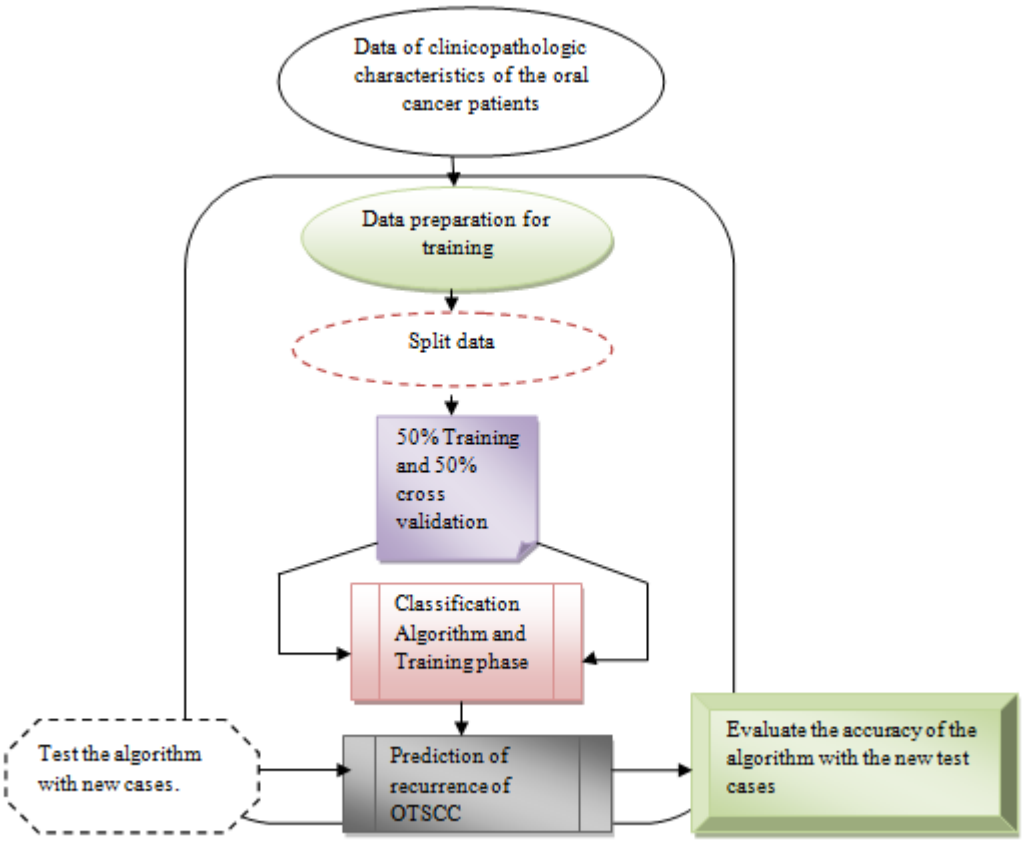

Figure 1

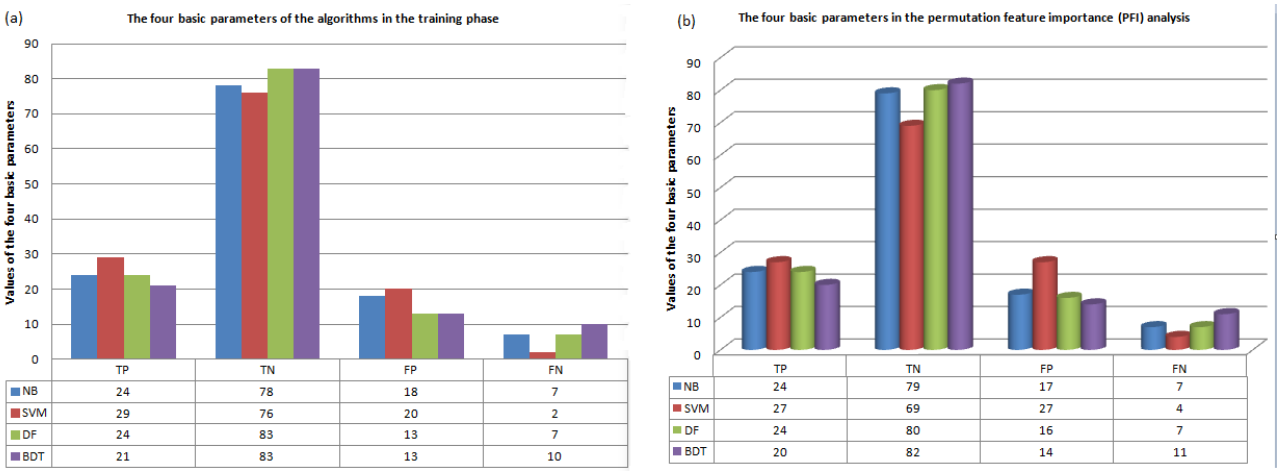

Figure 2

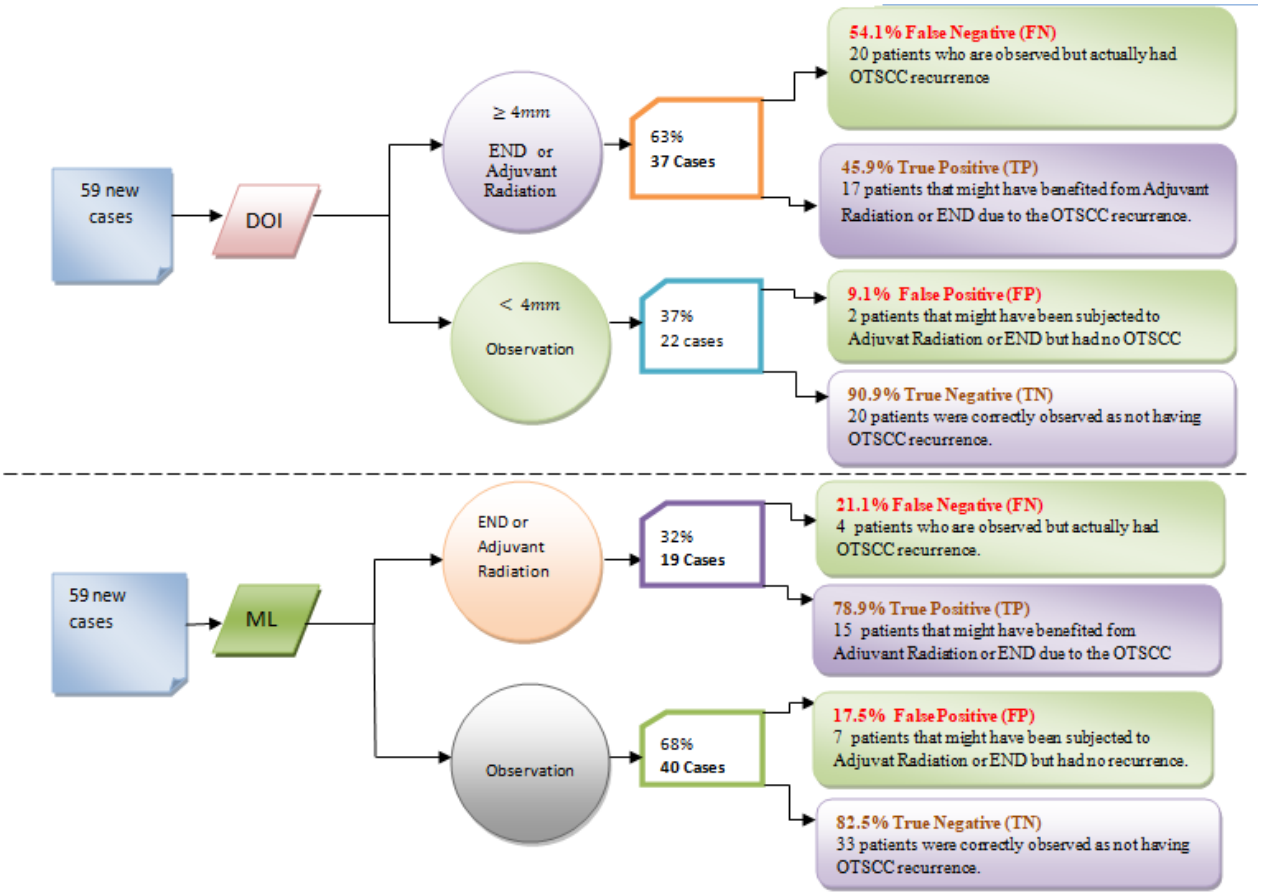
59 new cases
DOI
≥ 4mm
END or Adjuvant Radiation
63%
37 Cases
54.1% False Negative (FN)
20 patients who are observed but actually had OTSCC recurrence
45.9% True Positive (TP)
17 patients that might have benefited fom Adjuvant Radiation or END due to the OTSCC recurrence.
< 4mm
Observation
37%
22 cases
9.1% False Positive (FP)
2 patients that might have been subjected to Adjuvat Radiation or END but had no OTSCC
90.9% True Negative (TN)
20 patients were correctly observed as not having OTSCC recurrence.
59 new cases
ML
END or Adjuvant Radiation
32%
19 Cases
21.1% False Negative (FN)
4 patients who are observed but actually had OTSCC recurrence.
78.9% True Positive (TP)
15 patients that might have benefited fom Adjuvant Radiation or END due to the OTSCC
Observation
68%
40 Cases
17.5% False Positive (FP)
7 patients that might have been subjected to Adjuvat Radiation or END but had no recurrence.
82.5% True Negative (TN)
33 patients were correctly observed as not having OTSCC recurrence.

Figure 3